

Operation Analytics and Investigating Metric Spike

PROJECT DESCRIPTION:-

There are two case studies:- This project involves two case studies: operational analysis of job data and investigation of metric spikes in user engagement. The goal is to enhance operational efficiency and optimize user interactions with email services. Key metrics include throughput, language distribution, duplicate rows in job data, user engagement, user growth, weekly retention, and email engagement.

APPROACH:-

The project involves creating a database and tables to store the data. SQL is employed for advanced analytics and extracting insights. The analysis focuses on metrics such as job review rates, throughput fluctuations, language distribution, multilingual job activities, user engagement trends, user growth patterns, weekly retention, and email engagement metrics.

TECH-STACK USED:-

The tech stack for this project includes SQL is used for advanced analytics and extracting valuable insights from the datasets.

INSIGHTS:-

Case Study 1:- Operational Analysis of job data

1.Number of jobs reviewed:

QUERY:

```
SELECT DISTINCT ds AS days,  
Count(job_id) / (Sum(time_spent) / 3600) AS no_of_jobs_reviewed  
FROM job_data  
GROUP BY days;
```

days	no_of_jobs_reviewed
2020-11-29	180.0000
2020-11-28	218.1818
2020-11-27	34.6154
2020-11-30	180.0000
2020-11-26	64.2857
2020-11-25	80.0000

- On average, 126 jobs are reviewed per hour per day in November.
- The highest number of jobs was reviewed on 28 November 2020 with 218 jobs per hour.
- The lowest number of jobs was reviewed on 27 November with 34 jobs per hour.

2.Throughput:

QUERY:

```
SELECT a.ds AS day,
a.throughput,
avg(a.throughput) OVER ( ORDER BY ds rows BETWEEN 6 PRECEDING
AND CURRENT row ) AS
7_day_avg_of_throughput
FROM
( SELECT ds, count(job_id) / sum(time_spent) AS throughput FROM
job_data GROUP BY ds ) AS a
GROUP BY ds;
```

```
SELECT language,
count(job_id) as no_of_jobs,
count(job_id)*100 / sum(count(*)) OVER() as percentage_share
```

```
FROM job_data
WHERE ds between '2020-11-01' and '2020-11-30'
GROUP by language;
```

day	throughput	7_day_avg_of_throughput
2020-11-25	0.0222	0.02220000
2020-11-26	0.0179	0.02005000
2020-11-27	0.0096	0.01656667
2020-11-28	0.0606	0.02757500
2020-11-29	0.0500	0.03206000
2020-11-30	0.0500	0.03505000

- **Throughput Fluctuations:** The data show variations in the throughput values from day to day, indicating changes in system performance or workload.

- **Higher 7-Day Average Throughput:** The 7-day average throughput is consistently higher than the individual daily throughput values, suggesting an overall higher level of throughput during the given time period.

3. Percentage share of each language:

QUERY:

```
SELECT language,
count(job_id) as no_of_jobs,
count(job_id)*100 / sum(count(*)) OVER() as percentage_share

FROM job_data
WHERE ds between '2020-11-01' and '2020-11-30'
GROUP by language;
```

language	no_of_jobs	percentage_share
Persian	3	37.5000
Hindi	1	12.5000
French	1	12.5000
English	1	12.5000
Arabic	1	12.5000
Italian	1	12.5000

- **Persian Language Dominance:** Persian is the most prominent language in the dataset, representing 37.5% of the total jobs.
- **Equal Distribution of Other Languages:** Hindi, French, English, Arabic, and Italian each have an equal share of 12.5% of the total jobs, indicating a balanced distribution among these languages.
- **Multilingual Dataset:** The dataset showcases a diverse range of languages, highlighting the need for multilingual support in job-related contexts.

4. Duplicate rows:

QUERY:

```
SELECT a.ds,
a.job_id,
```

a.actor_id,

a.event,

a.language,

a.time_spent,

a.org,

CASE when a.duplicates = 1 then "No Duplicate" else "Duplicate" end as Duplicate

FROM

(SELECT *, row_number() OVER (partition by ds, job_id, actor_id, event, language, time_spent, org)as duplicates FROM job_data) as a ;

	ds	job_id	actor_id	event	language	time_spent	org	Duplicate
▶	2020-11-25	20	1003	transfer	Italian	45	C	No Duplicate
	2020-11-26	23	1004	skip	Persian	56	A	No Duplicate
	2020-11-27	11	1007	decision	French	104	D	No Duplicate
	2020-11-28	23	1005	transfer	Persian	22	D	No Duplicate
	2020-11-28	25	1002	decision	Hindi	11	B	No Duplicate
	2020-11-29	23	1003	decision	Persian	20	C	No Duplicate
	2020-11-30	21	1001	skip	English	15	A	No Duplicate
	2020-11-30	22	1006	transfer	Arabic	25	B	No Duplicate

- **Multilingual Job Activities:** The dataset reveals job activities involving multiple languages, including Italian, Persian, French, Hindi, English, and Arabic, reflecting a diverse linguistic environment.
- **No Duplicate Entries:** The dataset does not contain any duplicate entries, indicating that each combination of the date, job ID, actor ID, event, language, time spent, and organization is unique.

INSIGHTS:-

Case Study 2:- Investigating metric spike:

1.User Engagement:

QUERY:

```
SELECT week(occurred_at) as Week,  
count(DISTINCT user_id)as Weekly_User_engagement  
FROM events  
GROUP BY week(occurred_at)  
ORDER BY week(occurred_at);
```

Week	Weekly_User_engagement
17	740
18	1260
19	1287
20	1351
21	1299
22	1381
23	1446
24	1471
25	1459
26	1509
27	1573
28	1577
29	1607
30	1706
31	1514
32	1454
33	1438
34	1443
35	118

- **Increasing User Engagement:** The weekly user engagement data shows a generally increasing trend, starting from week 17 with a value of 740 and reaching its peak in week 32 with a value of 1706, indicating a growth in user activity and interaction over time.
- **Fluctuations in User Engagement:** Although there is an overall increasing trend, there are fluctuations in user engagement throughout the weeks. For example, there is a slight decrease in week 18 followed by a significant increase in week 19 and a subsequent dip in week 20. These fluctuations suggest variations in user behavior and engagement levels.
- **The weekly user engagement data demonstrates a rising trend in overall user activity, with some fluctuations observed throughout the weeks.**

2.User Growth:

QUERY:

```

SET @g := 0;
SELECT a.no_of_users, a.date,
( @g := @g + a.no_of_users ) as user_growth
FROM
( SELECT count(user_id) as no_of_users,
date(created_at) as date
FROM users WHERE state = "active"
GROUP BY date(created_at) ) a;

```

no_of_users	date	user_growth
7	2013-01-01	7
7	2013-01-02	14
6	2013-01-03	20
1	2013-01-04	21
2	2013-01-05	23
3	2013-01-06	26
4	2013-01-07	30
2	2013-01-08	32
6	2013-01-09	38
6	2013-01-10	44
6	2013-01-11	50
3	2013-01-12	53

- **User Growth Trend:** The data show consistent growth in the number of users over time, with an increase from 7 users on 2013-01-01 to 53 users on 2013-01-12. This indicates a positive trend in user acquisition and engagement.
- **Varied Daily User Growth:** The daily user growth varies throughout the period. For example, there is a significant increase in user growth from 7 to 14 on 2013-01-02, followed by a smaller growth of 6 on 2013-01-03. These fluctuations suggest that user acquisition may vary on a day-to-day basis.
- The data illustrates a consistent overall user growth trend, with variations in daily user acquisition, showcasing positive growth in the number of users over the given period.

3.Weekly Retention:

QUERY:

```
SELECT user_id,
activated_at
FROM users
WHERE activated_at > '2014-05-01'
ORDER BY user_id;
```

user_id	activated_at
11768.0	2014-05-01 08:03:12
11770.0	2014-05-01 06:08:50
11775.0	2014-05-01 16:38:06
11778.0	2014-05-01 18:49:49
11779.0	2014-05-01 18:24:54
11780.0	2014-05-01 10:34:04
11785.0	2014-05-01 07:20:37
11787.0	2014-05-01 18:23:18
11791.0	2014-05-01 15:50:33
11793.0	2014-05-01 09:29:48
11795.0	2014-05-01 03:44:18
11798.0	2014-05-01 23:12:34
11799.0	2014-05-01 12:07:10
11801.0	2014-05-01 10:15:50
11804.0	2014-05-01 08:09:58

```

SELECT DISTINCT u.user_id,
e.occurred_at
FROM users u join events e on u.user_id = e.user_id
WHERE u.activated_at > '2014-05-01' and e.event_name = 'login'
GROUP BY week( e.occurred_at)
ORDER BY e.occurred_at;

```

user_id	occurred_at
11768.0	2014-05-01 08:03:12
11775.0	2014-05-09 12:21:24
11787.0	2014-05-13 15:59:11
11799.0	2014-05-20 07:56:03
11778.0	2014-05-28 14:10:56
11901.0	2014-06-03 15:12:16
11778.0	2014-06-09 07:06:11
12741.0	2014-06-20 14:31:18
12882.0	2014-06-22 18:46:04
13441.0	2014-07-04 10:33:32
13942.0	2014-07-12 08:53:20
13757.0	2014-07-18 20:39:28
13317.0	2014-07-20 12:18:05
13743.0	2014-07-30 06:10:40
14227.0	2014-08-04 22:54:30

- **User Activation and Occurrence:** The data present user activation events and their corresponding occurrences. The user IDs indicate individual users, and the occurred at/activated timestamps show the dates and times of activation and occurrence.
- **User Activation Dates:** The user activation dates range from 2014-05-01 to 2014-05-01, indicating a single-day activation period for these users.
- **Multiple Occurrences:** Some user IDs, such as 11778.0, have multiple occurrences on different dates, suggesting repeated interactions or events by the same user.

- The data provides insights into user activation and occurrences, with a single-day activation period and instances of multiple occurrences for certain user IDs.

4.Weekly Engagement:

QUERY:

```
SELECT week(occurred_at) as Weeks,device,
count(distinct user_id)as User_engagement
FROM events
GROUP BY device,
week(occurred_at)
ORDER BY week(occurred_at);
```

Weeks	device	User_engagement
17	acer aspire desktop	12
17	acer aspire notebook	23
17	amazon fire phone	4
17	asus chromebook	23
17	dell inspiron desktop	20
17	dell inspiron notebook	48
17	hp pavilion desktop	17
17	htc one	19
17	ipad air	29
17	ipad mini	19
17	iphone 4s	27
17	iphone 5	69
17	iphone 5s	47
17	kindle fire	6
17	lenovo thinkpad	94
17	mac mini	7
17	macbook air	61
17	macbook pro	154

- **User Engagement by Device:** The data provides user engagement information for different devices, including Acer Aspire desktop, Acer Aspire notebook, Amazon Fire Phone, Asus Chromebook, Dell Inspiron desktop, Dell Inspiron notebook, HP Pavilion desktop, HTC One, Pad AR, Pad Mini, Phone 4S, Phone 5, iPhone 5S, Kridle Fire, Lenovo ThinkPad, MacMini, MacBook Air, and MacBook Pro.
- **Varied User Engagement Levels:** The user engagement values range from 4 to 154, indicating varying levels of engagement across different devices. For example, the highest user engagement of 154 is associated with the MacBook Pro.
- The data reveals user engagement levels for different devices, with varying levels of engagement observed, including the highest engagement with the MacBook Pro device.

5.Email Engagement:

QUERY:

```
SELECT week(occurred_at) as Week,  
count( DISTINCT ( CASE WHEN action = "sent_weekly_digest"  
THEN user_id end )) as weekly_digest,  
count( distinct ( CASE WHEN action = "sent_reengagement_email"  
THEN user_id end )) as reengagement_mail,  
count( distinct ( CASE WHEN action = "email_open"  
THEN user_id end )) as opened_email,  
count( distinct ( CASE WHEN action = "email_clickthrough"  
THEN user_id end )) as email_clickthrough  
FROM emails  
GROUP BY week(occurred_at)  
ORDER BY week(occurred_at);
```

Week	weekly_digest	reengagement_mail	opened_email	email_clickthrough
17	908	73	310	166
18	2602	157	900	425
19	2665	173	961	476
20	2733	191	989	501
21	2822	164	996	436
22	2911	192	965	478
23	3003	197	1057	529
24	3105	226	1136	549
25	3207	196	1084	524
26	3302	219	1149	550
27	3399	213	1207	613
28	3499	213	1228	594
29	3592	213	1201	583
30	3706	231	1363	625
31	3793	222	1338	444

- **Weekly Process:** The number of weekly process emails sent to clients ranged from 908 to 3793, indicating a potential growth in the user base or email subscribers over time.
- **Reengagement Mail:** The data revealed varying levels of engagement with re-engagement mail, including the number of opened emails and email clickthroughs, suggesting different levels of user interaction with the content.
- **Email Engagement:** User engagement with emails, measured by opened emails and email clickthroughs, fluctuated across weeks, indicating varying levels of user response and interaction with the email content.
- Overall, the findings suggest a potential increase in the user base, with varying levels of engagement observed for re-engagement mail and email content.

- **RESULT:-**

The project involved two case studies: operational analysis of job data and investigating metric spikes in user engagement. The approach included creating a database, using SQL for advanced analytics, and extracting insights. Key findings include an average of 126 jobs reviewed per hour in November, Persian being the dominant language with 37.5% share, no duplicate entries in the job data, increasing user engagement with fluctuations over time, consistent user growth, single-day user activation with some instances of multiple occurrences, varied user engagement levels across different devices, and increasing weekly digest emails indicating potential user base growth.